McKinzey, R.K (2005, November 28) Author's update to "Rorschach Interscorer Agreement". *WebPsychEmpiricist*. Retrieved (date) from: http://www.wpe.info/papers_table

WPE WebPsychEmpiricist

Author's Update to "Rorschach Interscorer Agreement"

R. K. McKinzey

Oakland, CA

11/28/05

Editor's Note: This update file comments upon events and research available after the publication of WPE's "The Rorschach, Exner's Comprehensive System, Interscorer Agreement, and Death"(McKinzey & Campagna, 2002, April 27) This is the fourth update. Before citing or using this article, check for later updates.

Abstract

Since the 5/1/02 publication of "The Rorschach, Exner's Comprehensive System, Interscorer Agreement, and Death", other relevant articles have become available and are reviewed. In the first update (2/22/03) an article admitted as true several criticisms made by the CS critics. Another, by Exner and colleagues, presented an interscorer reliability study using the same paired design as called for by the critics and as done by three previous studies. The results are so completely different that they must be viewed as suspect. A third article is supportive, but only through apparently selective reporting. In the second update (4/14/03), a new book presents the case against the CS in detail, citing our article. In this third update (6/16/05), two more articles are reviewed, both of which have cited the paper. One agrees with the misscoring problems, but insists they are due to human error, rather than the test itself. The second challenged the use of a dissertation (Mittman, 1983) in supporting a citation of the Rorschach's false positive rate at 81%. The fourth update (8/10/05) notes a Rorschach apologist has written about how easily a given Rorschach can be rebutted.

Correspondence regarding this article should be addressed to : R.K. McKinzey, Ph.D., 400 29th St., Ste. 315, Oakland, CA. 94609 510-655-3903, email: editor@wpe.info or WPEList@yahoogroups.com

Author's Update to "Rorschach Interscorer Agreement"

2/22/03 (later updates follow)

In 1996, Exner's Comprehensive System (CS) for the Rorschach was heavily criticized by James Wood and his colleagues (Wood, Nezworski, & Stejskal, 1996). Wood et al. argued that:

1. Exner had not properly established the interrater reliability of the CS. Instead of citing how often a pair of scorers agreed with each other on many protocols (a formal paired design using kappa), Exner had cited how often many scorers agreed with a standard of correctness, a study now refuted (Guarnaccia, Dill, Sabatino, & Southwick, 2001; McKinzey & Campagna, 2002, April 27).
2. Exner had wrongly denied Wood the access to experimental data so crucial to check the work of scientists.
3. Exner had not adequately described or published too many studies underlying the CS.
4. Much of Exner's work had failed to replicate, but Exner was slow to acknowledge such failures. Wood et al. cited as examples such CS variables as EGOI, D, DEPI, S-CON, and SCZI.

Wood et al. proved to be seminal. A long series of both review and experimental articles followed, including McKinzey & Campagna. Gregory Meyer, now editor at *Journal of Personality Assessment* (started by the Rorschach industry to ease publication of Rorschach research) has proven to be a capable defender of the Rorschach CS industry. He and colleagues have published two articles[1] that bear directly on McKinzey & Campagna.

The first article is the final one in a special series on the CS in *Psychological Assessment* (Meyer & Archer, 2001). In it, Meyer artfully[2] admits that:

1. Exner should have described his unpublished research better.
2. Exner should have been more cooperative with his potential critics.
3. DEPI and SCZI failed to replicate.
4. To properly establish the CS' interrater reliability, a formal paired design should be done.

---

[1] Editor's note: Request pdf reprints from Dr. Meyer using WPE's Reprints Available table at: http://wpe.info/reprints_available.html

[2] The article admits that the Wood et al.'s specific criticisms are correct without admitting or even addressing the more general issue, that, since so much of Exner's work has failed replication, all of his work must be reexamined.

The second article, Meyer, Hilsenroth, Baxter, Exner, Fowler, Piers, & Resnick (2002). is that reliability study, actually five separate studies. The first used 66 outpatient (clinical) protocols, 165 CS variables, and Meyer's scoring compared to different groups of student raters. 96% of the variables passed. The second used 65 clinical protocols, several experienced researchers, and 140 variables. 100% passed. The third used 19 clinical protocols, multiple clinicians, and 135 variables. 98% passed. The fourth used 69 inpatient protocols, clinicians and experienced researchers, and 139 variables. 100% passed. The fifth used a mix of the other four studies, 219 protocols, and 138 variables. 99% passed.

These numbers are startling better than the other three studies using the formal paired design and comprehensive surveys of CS variables. The first (Shaffer, Erdberg, & Haroian, 1999) had 41% fail in a normal sample. The second (Nakata, 1999) had 37% fail in the clinical sample, and 26% fail in the normal sample. The third (Acklin, McDowell, Verschell, & Chan, 2000) had 9% fail in the clinical sample and 20% in the normal sample.

Wood et al. asserted a fundamental rule of Science: If the results cannot be replicated by independent scientists, the results are suspect. To forestall this criticism being leveled at their too-good results, Meyer and Exner argue their "results are essentially equivalent to those observed by other investigators" (p. 267) and cite 18 other studies.

Their cites do not bear inspection. Of the 18, four (McDowell & Acklin, 1996; Meyer, 1997; Perry, McDougall, & Viglione, 1995a; Perry et al., 1995b) only reported "segments", which are kinds of variables, not the variables themselves, making the cites not comparable. Of the 14 left, only 2 are comprehensive—and are Acklin et al. and Shaffer et al., who had vastly different results! Nakata was not cited.

Of the other 12, two (Archer & Krishnamurthy, 1997; Krishnamurthy, Archer, & House, 1996) are different reports of the same study. These two and the remaining 10 studies (Baity & Hilsenroth, 1999; Franklin & Cornell, 1997; Greco & Cornell, 1992; Hilsenroth, Fowler, & Padawer, 1998; Netter & Viglione, 1994; Ornduff, Centeno, & Kelsey, 1999; Perry & Braff, 1994; Perry, Potterat, Auslander, Kaplan, & Jest, 1996; Perry & Viglione, 1991; Young, Justice, & Erdberg, 1999)[3] were validation studies, which examined new, non-CS variables in addition to 3-16 CS variables. All of these 11 studies reported 100% of their CS variables to pass interscoring reliability statistics.

[3] One of these many accomplished researchers is the administrating psychologist.

However, these studies tell us nothing of how the CS stands up to scrutiny. First, all of these studies only used selected CS variables. If these same researchers had used a larger set of CS variables (making the study comparable to the four comprehensive studies), they might have gotten entirely different results. Of the four comprehensive studies, only the one by Exner produced sterling results. Once again, Exner has offered results no one else can replicate.

In January 2003, another paired design reliability study (Viglione & Taylor, 2003)[4] was published in *Journal of Clinical Psychology* (see "Authors' Reply" for a discussion of how JCP handled our article) by Donald Viglione, who is a member of the Rorschach Research Council, a teacher for the Rorschach Workshops, Exner's former graduate student at LIU, and author of a new CS scoring book. This study used 84 mixed normal and clinical protocols and reported on 73 CS variables; only 3% failed. The authors did not report or calculate the full set of CS variables,[5] so the study cannot be said to either support Meyer et al. (2002) or to be comparable to the other reliability studies. While they criticize other authors for leaving out unsupportive studies in literature reviews, their own review did not mention Schaffer et al., Nakata, Guarnaccia et al., or McKinzey & Campagna.

In our article, we asked, "if the scorers can't agree on the scores, of what use is the test?" We note that the researchers, working in laboratory conditions, cannot agree on how many CS variables are reliably scored, which variables are unable to be scored, or even which variables ought to be reported. How, then, can we be sure that individual scorers in the field are scoring this evidently difficult test at acceptable levels? This question, like all the others we asked in our article, remain unanswered.

At the moment, the Rorschach industry is in a quandary. They and the critics have found that the norms and some variables are useless. The critics now argue that the entire literature is in shambles and must be abandoned. The industry, loath to desert so many hours of research, teaching, and patient protocols, are still trying to establish what can be saved. The norms are being redone and efforts made to increase scoring accuracy. However, it will be years while the Rorschach CS literature is rehabilitated, if at all. In the meantime, we ask, "Why would anyone still use such a test? Or pay for it?"

---

[4] Editor's note: Request a pdf reprint from Dr. Viglione using WPE's Reprints Available table at: http://wpe.info/reprints_available.html

[5] D. Viglione, personal communication, 4/5/03.

4/14/03

The controversy over the Rorschach started long before the CS was devised. A new book, *What's Wrong With the Rorschach* (Wood, Nezworski, Lilienfeld, & Garb, 2003) lucidly reviews the history of the controversy—and then shows how the same issues are still bedeviling it 80 years after the Rorschach's introduction. One of those long-standing issues, of course, is insufficient levels of interscoring reliability.

McKinzey & Campagna is cited once, and the case is described elsewhere. Wood et al. (2003) also review the CS reliability studies (pages 227-234: q.v. footnote 59, page 366), and argue that the sterling pass rates obtained by the CS defenders are simply the result of mixing samples, a statistical design mistake. When samples with different sets of judges are combined, the resulting ICC statistic can be inflated.

The book can be bought from Amazon via WPE's Notable Books feature at: http://wpe.info/books.html

6/16/05

McKinzey & Campagna has been cited twice (Hilsenroth & Stricker, 2004; McGrath, 2003) for the first time in journal articles.

In defending the CS, the Rorschach industry has taken a "et tu" stance. That is, they argue that the CS has no problems other tests don't have, and any attempts to discontinue its use are unfair, biased attacks. In a review article, McGrath argues that many tests are routinely misscored, citing a string of articles showing misscoring on the WAIS family of IQ tests. He then notes that similar research exists for the Rorschach CS, citing three studies (Exner, 1988; Guarnaccia et al., 2001; McKinzey & Campagna, 2002, April 27).

Exner sent out quizzes to alumni of various levels of CS workshops. One quiz had 23 responses to be scored, the other 17. "Only 'major' errors were recorded for the determinants." (p. 4) Only means (& percentages) or errors *per protocol* for each workshop level and variable category were reported, without either ranges, SD, or the percentages of total errors per protocol. For DQ, the scorers averaged about 17% errors; for Determinants, the scorers averaged about 15% errors; for Z Scores, the scorers averaged about 9% errors,; and for Special Scores, the scorers averaged about 29% errors. "The overall results are disconcerting" (p. 5) and "significantly" worse than previous "interscorer reliability studies done with other groups and in house staff…If the bulk of the interpretation is generated from a Structural Summary that has

average error rates similar to those in Table 1, the results would be misleading, and even totally wrong in some cases." (p. 5) The article then offers more precise scoring instructions.

Exner (1988) was noted by Wood et. al (1996). In his reply article (Exner, 1996) he explained that he thought he had eliminated the problem after tweaking the CS, having done a similar, uncited study in 1994.

McGrath then comments on McKinzey & Campagna: "Although the study was methodologically[6] flawed, one finding was unequivocally relevant to the issue of scoring accuracy…19 of 27 scorers (70%) made errors." (p. 105) However, we "made the error of blaming these problems on the test rather than the tester…The risk of testing inaccuracies is inevitable, and there is probably no way to eliminate the problem completely short of removing the clinician from the testing process. However, there are ways to minimize the risk." (p. 108) He recommends having two scorers, and use of a credentialing system. He does not mention the possibility of using a test that has been independently demonstrated to be possible to be scored correctly.

It may be that tests involving the transformation of verbal responses to numbers need to be more closely examined. The Rorschach and WAIS are certainly examples, but there may also be more. Although even objective tests like the MMPI *can* be misscored, there is no evidence they *cannot* be scored accurately. So far, the <u>published</u> literature indicates the same cannot be said of the CS. If there are <u>unpublished</u> studies showing it *can* be scored accurately, they should be presented. Whether the Rorschach CS industry joins McGrath in blaming the users of the test rather than the test itself remains to be seen.

Hilsenroth & Stricker[7] address the problem of defending the Rorschach in court. First, they recommend "knowing the current research literature…It is important not to accept any article, chapter, or book as a final authority if a psychologist can imagine disagreeing with any of the findings or conclusions in some circumstances or characterizations." (p. 141). They then cite the supportive studies without mentioning the unsupportive ones, and reassert the et tu argument that the Rorschach is no worse than other tests. They note that the test is still being renormed, that the DEPI should not be used, that the test should not be used to diagnose sexual abuse, &

---

[6] For an answer to the charge of McKinzey & Campagna being "methodologically flawed", see our Author's Reply and Footnote 3 of this Update.

[7] This article has been added to WPE's Reprints Available page.

shouldn't be used alone. They argue that the test will meet courtroom admissibility standards[8]. They note that forensic psychologists have reported not using the Rorschach (Lally, 2003), commenting "The Rorschach may not have been a good choice…but it is puzzling why the other tests were considered to be better." (p 146)

They then argue that the debate has become adversarial, with "More stringent standards seem to be used to evaluate studies that report positive Rorschach results than negative ones, and negative findings are often cited, whereas positive results in the same study are not." (p. 147) As an example "of an incomplete presentation of data", they cite two McKinzey articles (Brodsky & McKinzey, 2002; McKinzey & Campagna, 2002, April 27). Brodsky & McKinzey had cited McKinzey & Campagna's death penalty case, and noted the CS' 81% false positive rate (Mittman, 1983), "discredited norms" (p. 148) (Wood, Nezworski, Garb, & Lillienfeld, 2001), and questionable scoring system (Guarnaccia et al., 2001; Wood et al., 1996).

They challenge the use of Mittman's dissertation, apparently without having read it: Mittman, they protest, "concerned whether individuals could malinger schizophrenia and not the implied validity" (p. 148) of the CS. Since the use of Mittman has been challenged, a review of it has been written for WPE (McKinzey, 2005, May 15), describing how it found an 81% false positive rate.

They then protest that the norms & scoring system are not troublesome!

"It is surprising that such emphasis was placed on an unpublished manuscript…that had been rejected from two different journals" (p. 148), citing WPE's publication of the peer reviews (McKinzey, 2002, April 30) without citing the response to those reviews (McKinzey & Campagna, 2002, April 30).

They conclude that any testifying psychologist should "be aware of original sources and the current research literature. Only then can the use of inaccurate and misleading information be exposed for the lack of comprehensive scholarship it represents." (p. 148) They also reiterate that no single CS variable should be used in isolation.

They never address the sole conclusion of McKinzey & Campagna: The Rorschach should not be used in court simply because it is easily rebutted. A psychologist testifying using it runs the risk of having it ridiculed, rescored, declared invalid or faked, or challenged on

[8] In reply to this argument, Wood et al. (2003) cited McKinzey & Ziegler (1999).

interpretation by the rebutting psychologist, not to mention being confronted with the literature on cross-examination. We ask again: Why would anyone use such a test? Or pay for it?

8/10/05

We concluded that the Rorschach is easily rebutted. One of the methods to rebut a given protocol's interpretation is to simply rescore it. One Rorschach apologist, Carl B. Gacono, describes a case in which he did just that, simply by challenging the inquiry (Gacono, Evans, & Viglione, 2002).

References

Acklin, M. W., McDowell, C. J., Verschell, M. S., & Chan, D. (2000). Interobserver agreement, intraobserver reliability, and the Rorschach Comprehensive System. *Journal of Personality Assessment, 74*(1), 15-47.

Archer, R. P., & Krishnamurthy, R. (1997). MMPI-A and Rorschach indices related to depression and conduct disorder: An evaluation of the incremental validity hypothesis. *Journal of Personality Assessment, 69*(3), 517-533.

Baity, M. R., & Hilsenroth, M. J. (1999). Rorschach aggression variables: A study of reliability and validity. *Journal of Personality Assessment, 72*(1), 93-110.

Brodsky, S. L., & McKinzey, R. K. (2002). The Ethical Confrontation of the Unethical Forensic Colleague. *Professional Psychology: Research & Practice, 33*(3), 307-309.

Exner, J. E. (1988). Scoring issues. *Alumni Newsletter*, 4-8.

Exner, J. E., Jr. (1996). A comment on "The Comprehensive System for the Rorschach: A critical examination. *Psychological Science, 7*(1), 11-13.

Franklin, K. W., & Cornell, D. G. (1997). Rorschach interpretation with high-ability adolescent females: Psychopathology or creative thinking? *Journal of Personality Assessment, 68*(1), 184-196.

Gacono, C. B., Evans, F. B., & Viglione, D. J. (2002). The Rorschach in forensic practice. *Journal of Forensic Psychology Practice, Vol 2*(3), 33-54.

Greco, C. M., & Cornell, D. G. (1992). Rorschach object relations of adolescents who committed homicide. *Journal of Personality Assessment, 59*(3), 574-583.

Guarnaccia, V., Dill, C. A., Sabatino, S., & Southwick, S. (2001). Scoring accuracy using the Comprehensive System for the Rorschach. *Journal of Personality Assessment, 77*(3), 464-474.

Hilsenroth, M. J., Fowler, J. C., & Padawer, J. R. (1998). The Rorschach Schizophrenia Index (SCZI): An examination of reliability, validity, and diagnostic efficiency. *Journal of Personality Assessment, 70*(3), 514-534.

Hilsenroth, M. J., & Stricker, G. (2004). A consideration of challenges to psychological assessment instruments used in forensic settings: Rorschach as exemplar. *Journal of Personality Assessment, 83*(2), 141.

Krishnamurthy, R., Archer, R. P., & House, J. J. (1996). The MMPI-A and Rorschach: A failure to establish convergent validity. *Assessment, 3*(2), 179-191.

Lally, S. J. (2003). What tests are acceptable for use in forensic evaluations? A survey of experts. *Professional Psychology: Research & Practice, 34*(5), 491-498.

McDowell, C., & Acklin, M. W. (1996). Standardizing procedures for calculating Rorschach interrater reliability: Conceptual and empirical foundations. *Journal of Personality Assessment, 66*(2), 308-320.

McGrath, R. E. (2003). Enhancing accuracy in observational test scoring: The Comprehensive System as a case example. *Journal of Personality Assessment, 81*(2), 104.

McKinzey, R. K. (2002, April 30). Prior peer review of "Rorschach interscorer agreement". *WebPsychEmpiricist*. Retrieved April 30, 2002, from http://wpe.info/papers_table.html

McKinzey, R. K. (2005, May 15). The Rorschach's false positive rate is 81%. *WebPsychEmpiricist*. Retrieved May 15, 2005, from http://wpe.info/papers_table.html

McKinzey, R. K., & Campagna, V. (2002, April 27). The Rorschach, Exner's Comprehensive System, Interscorer Agreement, and Death. *WebPsychEmpiricist*. Retrieved April 27, 2002, from http://www.wpe.info/papers_table.html

McKinzey, R. K., & Campagna, V. (2002, April 30). Authors' Reply to prior review of "Rorschach interscorer agreement". *WebPsychEmpiricist*. Retrieved April 30, 2002, from http://www.wpe.info/papers_table.html

McKinzey, R. K., & Ziegler, T. (1999). Challenging a flexible neuropsychological battery under Kelly/Frye: A case study. *Behavioral Sciences & the Law, 17*, 543-551.

Meyer, G. J. (1997). On the integration of personality assessment methods: The Rorschach and MMPI. *Journal of Personality Assessment, 68*(2), 297-330.

Meyer, G. J., & Archer, R. P. (2001). The hard science of Rorschach research: What do we know and where do we go? *Psychological Assessment, 13*(4), 486-502.

Meyer, G. J., Hilsenroth, M. J., Baxter, D., Exner, J. E., Jr., Fowler, J. C., Piers, C. C., et al. (2002). An examination of interrater reliability for scoring the Rorschach comprehensive system in eight data sets. *Journal of Personality Assessment, 78*(2), 219-274.

Mittman, B. L. (1983). Judges' ability to diagnose schizophrenia on the Rorschach: The effect of malingering. *Dissertation Abstracts International*, (UMI No. 8325540)

Nakata, L. M. (1999). Interrater reliability and the Comprehensive System for the Rorschach: Clinical and non-clinical protocols. *Dissertation Abstracts International*, (UMI No. AAI9944405)

Netter, B. E. C., & Viglione, D. J. (1994). An empirical study of malingering schizophrenia on the Rorschach. *Journal of Personality Assessment, 62*(1), 45-57.

Ornduff, S. R., Centeno, L., & Kelsey, R. M. (1999). Rorschach assessment of malevolence in sexually abused girls. *Journal of Personality Assessment, 73*(1), 100-109.

Perry, W., & Braff, D. L. (1994). Information-processing deficits and thought disorder in schizophrenia. *American Journal of Psychiatry, 151*(3), 363-367.

Perry, W., McDougall, A., & Viglione, D. (1995a). A five-year follow-up on the temporal stability of the Ego Impairment Index. *Journal of Personality Assessment, 64*(1), 112-118.

Perry, W., Potterat, E., Auslander, L., Kaplan, E., & Jest, D. (1996). A neuropsychological approach to the Rorschach in patients with dementia of the Alzheimer type. *Assessment, 3*(3), 351-363.

Perry, W., Sprock, J., Schaible, D., McDougall, A., Minassian, A., Jenkins, M., et al. (1995b). Amphetamine on Rorschach measures in normal subjects. *Journal of Personality Assessment, 64*(3), 456-465.

Perry, W., & Viglione, D. J. (1991). The Ego Impairment Index as a predictor of outcome in melancholic depressed patients treated with tricyclic antidepressants. *Journal of Personality Assessment, 56*(3), 487-501.

Shaffer, T. W., Erdberg, P., & Haroian, J. (1999). Current nonpatient data for the Rorschach, WAIS--R, and MMPI-2. *Journal of Personality Assessment, 73*(2), 305-316.

Viglione, D. J., & Taylor, N. (2003). Empirical Support for interrater reliability of Rorschach Comprehensive System coding. *Journal of Clinical Psychology, 59*(1), 111-121.

Wood, J. M., Nezworski, M. T., Garb, H. N., & Lillienfeld, S. O. (2001). The misperception of psychopathology: Problems with the norms of the Comprehensive System for the Rorschach. *Clinical Psychology: Science and Practice, 8*(3), 350-373.

Wood, J. M., Nezworski, M. T., Lilienfeld, S. O., & Garb, H. N. (2003). *What's Wrong With the Rorschach? Science Confronts the Controversial Inkblot Test*. New York: Wiley.

Wood, J. M., Nezworski, M. T., & Stejskal, W. (1996). The Comprehensive System for the Rorschach: A critical examination. *Psychological Science, 7*, 3-10.

Young, M. H., Justice, J., & Erdberg, P. (1999). Risk factors for violent behavior among incarcerated male psychiatric patients: A multimethod approach. *Assessment, 6*(3), 243-258.